A Novel Approach in Data Mining for Representative Pattern Sets

*K.Kavitha[1], C.Anand[2]
Department of Computer Applications, Jeppiaar Engineering College, Chennai 600119, Tamil Nadu, India
*Corresponding author: Email: kavivenkatpavi@gmail.com

ABSTRACT

Frequent pattern mining often produces an enormous number of frequent patterns, which shows a great issues on viewing, understanding and further analysis of the derived patterns. This makes for finding a small number of representative patterns to best approximate all other patterns. To find a minimum representative pattern set with error free an algorithm called MinRP set is created.MinRPset produces the small solution that we can possibly have in practice under the given problem and it takes a reasonable amount of time to finish when the number of frequent closed patterns is below one million. MinRPset is very memoryspace-consuming and time-consuming on some dense datasets when the number of frequent closed pattern is more. To solve this problem, we propose another algorithm called FlexRPset, which uses one extra parameter K to allow users to make a trade-off between result size and efficiency. We adopt an additive approach to let the users make the trade-off satisfaction. Our experiment results show that MinRPset and FlexRPset produce fewer representative patterns than RPlocal—an efficient algorithm that is developed for solving the same problem.

Keywords: Frequent closed pattern, representative pattern set.

INTRODUCTION

Mining frequent patterns from several patterns is one of the most important concepts in data mining. Other data mining concepts can be derived from this concepts. It is the beginning of the data mining technical training because it gives the effective idea about data mining which is not extremely technical.

Pattern Sets: A pattern is a template, form or model which is used to create or to generate parts of things. In data mining we say that a pattern is a particular behavior of data, arrangement or formation that might be of a business interest. A frequent pattern sets are item sets, subsequences, or substructures that appears in a data set in a frequency manner with no less than a user specified threshold. A substructure can refer to a different structural forms such as sub graphs, sub trees or sub lattices which may be combined with item sets. If a substructure is produced frequently in a database is called a frequent pattern. Finding frequent patterns plays an important role in mining associations, correlations and many other interesting relationships among data. Moreover it is useful in data indexing, classifying, clustering and other data mining tasks.

Frequent item sets: Another concept is a frequent item set which is a type of pattern set. A frequent item set is a parameter that is specified by the user in the database. The parameter is called as a support of an item set. Every subset of a frequent item set is also a frequent pattern. This property is also called as Apriori property or downward closure property. It explains that we do not need to find a count of the item set if subset is not frequent. This will become possible because of the anti- monotone property of support. Frequent item sets should satisfy the minimum support of user threshold (Agrawal, 1993). The support for an item sets never exceeds support for a subset. If we divide the entire database in several partitions then an item set can be frequent only if it is frequent in at least one partition. To find a frequent item set we should go through all sub item sets which themselves are frequent due to the Downward Closure property. The frequent itemsets are found (N.Pasquier, 1999) to reduce the problem of association rules.

The prefix tree structure (Grahne and Zhu, 2003) is used to implement the method for finding frequent itemsets.Another method for storing and querying (Liu, Lu, Yu, 2007) the information of frequent item sets is compact disk based structure. In the real life databases, there are several thousands of association rules that can cause redundant data. Redundant data can be minimized (Pasquier, Bastide, 2000) by using the frequent item sets.

A high level overview of frequent pattern mining methods. Methods includes efficient and scalable methods for mining frequent patterns, mining interesting frequent patterns, impact to data analysis and mining applications, applications of frequent patterns and research directions.

System architecture: Finding a representative pattern set is the best approach for finding best approximate results. The representative pattern set is used in the local dataset.

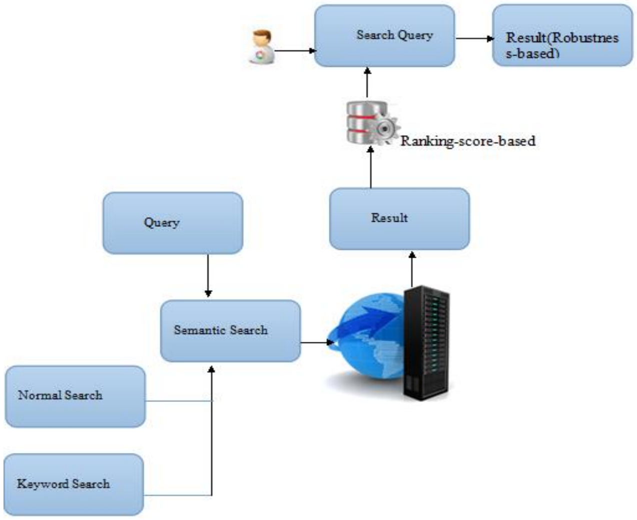

Fig1.Architecture diagram

Users first search the query in the local dataset. Local dataset contains a frequently used data by many users. The frequently used data generates the frequent pattern sets. These pattern sets are stored in the local dataset. Local dataset can retrieve the data from the dataset in the minimum amount of time. If the users not found a data from the local dataset the searching process is transferred to global dataset. Global dataset is large amount of space. The users can give the suggestions for the data used. Based on the suggestions the rank is provided to the data. Local dataset retrieves the best approximate result based on the frequent pattern set. Users can also search based on the ranking search.

Techniques: Frequent pattern mining is done based on data mining concept. This paper is describing about the finding representative pattern set. Finding the representative pattern set which describes the least number of sets that cover all frequent patterns. The problem of the MinRP set algorithm is to find the set of frequent patterns that covers over a large number of frequent pattern sets. To increase the efficiency of the algorithm following techniques can be used. They are considering closed patterns only, using a CFP tree structure and using a light weight compression techniques.

A. Considering closed patterns: A pattern is said to be closed if pattern is more frequent than its supersets.

B. Using CFP tree structure to find effective representative pattern sets: CFP tree structure is used to store the frequent pattern sets. To generate a frequent pattern set search CX algorithm is used. The frequent closed pattern set can be found by using DFS search CXs algorithm.

C. Compressing a frequent pattern sets: Finding a frequent closed pattern can produce a large number of frequent pattern sets. It makes the algorithm performance slow. To overcome this problem we compress a large number of frequent closed pattern set by using compression technique. MinRP set algorithm is used to compress the large amount of frequent pattern sets by removing non closed frequent pattern sets.

D. FlexRP set generates pattern sets: Searching a subset of frequent closed pattern sets makes the performance of MinRP set slow. It also consumes more memory space. To solve this problem we propose an algorithm called FlexRP set. This is used to find selective representative pattern set in a minimum number of times that covers all frequent closed pattern sets. It reduces time consuming and space consuming.

SYSTEM DESIGN

A. DataSet Training: To train the dataset to the search based on the pattern to retrieve a collection of documents related to the query pattern .After a query is submitted to a search engine, a list of Websnippets is returned to the user. We assume that if a keyword/phrase exists frequently in the Web-snippets of a specified query, it represents an important concept related to the query because it coexists in close proximity with the query in the top documents. Thus, we employ the following support formula, which is used for the well-known problem of finding frequent item sets in data mining measure the interestingness of a particular keyword/phrase ci extracted from the Web-snippets

B. Pattern Generation: All words from the text are not considered as Pattern. Usually some words occur frequently in almost all of the documents. Because of this property, their discrimination power is negligible. These types of words are called stop words and these words can be filtered out during patternization. Different types of stop word list [Probable Google’s, Onix 1,Onix 2,WordNet and RDS] have been tried in group, sole and absence, with the small update in global information of our system’s implementation.

C. Min RP Set: Matching is the process of generating the list of documents that match the query terms. We’ve implemented two types of matchers:-Match any query token – Documents that contain any term from the query are included in the matched list. We now review our personalized concept-based clustering algorithm with which ambiguous queries can be classified into different query clusters. Concept-based user profiles are generated in the clustering process to achieve personalization effect. First, a query-concept bipartite graph G is constructed by the clustering algorithm in which one set of nodes corresponds to the set of users’ queries and the other corresponds to the sets of extracted concepts. Each individual query submitted by each user is treated as an individual node in the bipartite graph by labeling each query with a user identifier.

D. Flex RP Set: Match all query tokens – Only documents that contain all the query tokens are included in the matched list. We defined co-occurrence measures using page counts. We showed how to extract clusters of patterns from snippets to represent numerous semantic relations that exist between two words.

CONCLUSION

A large number of frequent pattern sets is used to find the frequent closed pattern sets. A frequent closed pattern generates a representative pattern sets. For finding a representative pattern set two algorirhms MinRP set and FlexRP set are developed.MinRP set and Flexors set is used to keep record of the frequent closed pattern set which provides the best approximate result. MinRP set becomes slow when large amount of frequent closed pattern sets that consumes more memory space. MinRP set is expensive than RPlocal set. To overcome this problem, FlexRP set is proposed. FlexRP set adds extra parameter value to find the minimum number of representative pattern sets by covering all frequent closed pattern sets**.**

REFERENCES

A.Bykowski and C.Rigotti, A condensed representation to find frequent patterns, NewYork, NY, USA, 2001, 267-273.

A.K.Poernomo and V.GopalKrishnan,CP-summary: A concise representation for browsing frequent item sets, New York, NY, USA, 2009,687-696.

B.Goethals and M.J.Zaki Advances in Frequent itemset mining implementations: Introduction to FIMI’03, 2003.

C.Wang and S.Parthasarathy, Summarizing item set patterns using probabilistic models, Philadelphia, PA, USA, 2006, 730-735.

D.Xin, J.Han, X.Yan andH.Cheng, Mining compressed frequent-pattern sets, Trondheim, Norway, 2005, 709-720.

D.Xin,H.Cheng,X.Yan and J.Han, Extracting redundancy-aware top-k patterns, Philadelphia, PA, USA, 2006, 444-453.

F.N. Afrati, A.Gionis, and H.Mannila, Approximating a collection of frequent sets, Washington, DC, USA, 2004, 12-19.

G.Grahne and J.Zhu, Efficiently using prefix-trees in mining frequent itemsets, 2003.

G.Liu, H.Lu, and J.X.Yu, CFP-tree: A compact disk-based structure for storing and querying frequent itemsets, 295-319, 2007.

J.-FBoulicaut, A.Bykowski, and C.Rigotti, Free-sets: A condensed representation of Boolean data for the approximation of frequency queries, 5-22, 2003.

J.Pei,G.Dong,W.Zou, and J.Han, Mining condensed frequent bases,570-594,2004.

J.Wang,J.Han,Y.Lu and P.Tzvetkov, TFP: An efficient algorithm for mining top-k frequent closed itemsets,652-664.2005.

N.Pasquier, Y.Bastide, R.Taouil and L.Lakhal Discovering frequent closed itemsets for association rules, Jerusalem, Israel, 1999, 398-416

R.Agrawal,T.Imielinski, A.N.Swami, Mining association rules between sets of items in large databases Washington, USA 1993, 207-216.

R.J.Bayardo, Efficiently mining long patterns from databases, NewYork, NY, USA, 1998, 85-93.

R.Jin,M.Abu,Y.Xiang, and N.Ruan, Effective and efficient itemset pattern summarization: Regression-based approaches,Las Vegas,NV,USA,2008,399-407.

T.Calders and B.Goethals,Mining all non-derivable frequent itemsets,Helsinki, Finland,2002,74-85.

V.Chvatal, A greedy heuristic for the set-covering problem, 233-235, 1979.

X.Yan,H.Cheng,J.Han and D.Xin, Summarizing itemset patterns: A profile-based approach,Chicago,USA,2005,314-323.

Y.Bastide,N.Pasquier,R.Taouil,G.Stumme, and L.Lakhal, Mining minimal non-redundant association rules using frequent closed itemsets,London,U.K., 2000, 972-986.